DOCUMENT RESUME

ED 137 318 95 TM 005 930

AUTHOR Barker, Pierce; Pelavin, Sol H.
TITLE Issues of Reliability and Directional Bias in Standardized Achievement Tests: The Case of Mat70. P-5689.
SPONS AGENCY National Inst. of Educa[illegible] (DHEW), Washington, D.C.
PUB DATE Jul 76
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
AVAILABLE FROM The Rand Corporation, 1700 Main St., Santa Monica, California 90406 ($3.00)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Achievement Gains; *Achievement Tests; Educational Disadvantagement; Elementary Education; Grade Equivalent Scores; Program Evaluation; Raw Scores; Scores; Standard Error of Measurement; *Standardized Tests; Statistical Analysis; Test Bias; *Testing Problems; Test Interpretation; *Test Reliability; Test Validity
IDENTIFIERS Educational Voucher Demonstration; *Metropolitan Achievement Tests; Out of Level Testing; Standard Scores

ABSTRACT

This study was mounted to assess the validity of standard score transformations of raw test scores and test bias on the 1970 edition of the Metropolitan Achievement Test Battery, in the context of a controversial federally funded compensatory education program, the Educational Voucher Demonstration (EVD). On an individual level the validity of the Standard Score scale has not been demonstrated. Moreover, substantial bias was found in aggregate measures (means and medians) between adjacent difficulty levels. For these reasons, the authors could not conclude with any confidence that the instrument herein assessed could provide dependable bases either for individual student assessment, or program evaluation as usually performed. (HV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

AERA '76

ED137318

ISSUES OF RELIABILITY AND DIRECTIONAL BIAS IN STANDARDIZED ACHIEVEMENT TESTS: THE CASE OF MAT70

Pierce Barker

Sol H. Pelavin

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

July 1976

TM005 930

$ 300

The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

ISSUES OF RELIABILITY AND DIRECTIONAL BIAS IN STANDARDIZED ACHIEVEMENT TESTS: THE CASE OF MAT70

Pierce Barker

Harvard University

Sol H. Pelavin

Stanford Research Institute

Presented at the Annual Meeting of the
American Educational Research Association, 1976

INTRODUCTION

Standardized achievement tests (SAT's) for use in elementary and, to a perhaps lesser extent, secondary school grades, must perforce be designed to accommodate a very large degree of heterogeneity of actual performance level among students, both cross-sectionally and longitudinally.

The classical solution to this problem is twofold: within a given domain of knowledge, e.g., vocabulary, a series of tests is constructed, each of which is thought appropriate to a restricted age/grade range, in the sense that the 'average' student in a given range should score somewhere near the middle of the raw score distribution. This step is thought to be clearly necessary if the problem of excessive numbers of uninterpretably low and high scores is to be avoided, while keeping the test length within reasonable bounds.

However, for reasons explored in detail elsewhere (Barker and Pelavin, 1975, e.g.), simple aggregations or measures of central tendency computed on *raw scores* obtained from different test levels within a series are clearly inappropriate. Given the necessity to compute these measures, whether because a given nominal grade or classroom contains students achieving at widely different levels, or because one wishes to follow student cohorts through time, some form of transformation of the raw scores in the series is clearly required. This generally takes the form of what we will here call a mapping function, such that raw scores throughout the range on each of the successive levels are mapped upon a common baseline, which is taken to be a linear continuum of equal interval units. The assumptions and details of the most widely used technique are discussed in a number of standard sources, (Thurstone, 1925; Angoff, 1971; Guilford, 1954; Gulliksen, 1950, e.g.), and need not be rehearsed here. (For a fuller discussion in the present context, see Barker and Pelavin, 1975; a more generalized discussion of the problem appears in Porter and Chibucos, 1975.)

In general, then, the scores derived from the mapping function may be considered the basic metric of the system of measurement comprised by

the various SAT's. For the particular SAT system upon which this research is based, the Metropolitan Achievement Test battery, 1970 edition (Harcourt Brace Jovanovich, Inc., 1972, 1973), this metric is called the scale of Standard Scores (SS).[2] Now, it should be clear that, if this metric performs as intended, aggregation of scores derived from raw scores by the mapping function *is* appropriate; in fact, it may not be too much to argue that the intent of the mapping function is to transform the several levels of the various domain tests into parallel tests in the SS metric. (Gulliksen, 1950.)

That is, if one administers, say, two adjacent test levels within a single domain to a sample of students under suitable conditions, it is arguable that the correlation between transformed scores (in this case, SS) may be interpreted as a measure of the reliability of the tests under the assumptions underlying the theory of parallel tests. However, it seems more appropriate (as we argue at some length in Barker and Pelavin, 1975) to interpret such a correlation as a measure of the goodness of fit of the obtained transformed data to the hypothesis that they are parallel, i.e., as an investigation of the validity of that transformation. This is the approach adopted in the present research (see also Pelavin and Barker, forthcoming; 1976).

THE RESEARCH SETTING

Although an investigation of this sort would seem to have a certain amount of theoretical interest, the present research was in fact motivated by a keen concern about the validity of the transformations in the specific context of a federally funded and controversial educational intervention program for what we may, in admittedly crude shorthand, call educationally disadvantaged students in San Jose, California. This program, originally officially (and still most widely) known as the Educational Voucher Demonstration (EVD), is exhaustively described in various sources (e.g., Weiler et al., 1974; Weiner and Kellen, 1974), and the ethnic composition and SES of the student population--characteristics generally considered valid indicators of the degree of "educational disadvantage"--is described in, among other sources, Barker (1974).

Suffice here then to say that, while accelerated gains in measured cognitive achievement are neither direct nor immediate theoretical goals of the system of education vouchers as originally conceived (Jencks, et al., 1970; Barker, 1975(a)), the mandate to measure cognitive achievement delivered to the external evaluation staff suggested strongly that cognitive outcomes so measured might well have important policy implications; whereupon the dependability of conclusions drawn from these measures became an immediate and pressing issue.[3]

Moreover, this issue was not restricted to the problem of longitudinal analysis. Without question, in part as a result of the educational reorganization of schools in the EVD, such that the intended primary educational delivery units became relatively small, teacher-originated and teacher-directed programs (called, in the EVD, minischools), which also usually involved a high incidence of multigraded classrooms, nominal grades, minischools and classrooms did indeed contain students whose actual performance levels covered a wide range of test levels within a given domain. Consequently, the propriety of aggregation and computation of the moments of the distributions of test scores was a salient issue even for a single testing period, e.g., for the Fall of a given year. (A detailed report of the magnitude of what we may call the out-of-level testing problem appears in Barker and Pelavin, 1975.)

DESCRIPTION OF THE STUDY

Initially, we intended to mount a validation study which included a wide range of test levels and nominal grades; however, administrative strictures laid down by the local school district served to reduce the scope of the present study to one nominal grade and two adjacent test levels.[4]

During the regular Fall achievement testing period in 1973, all third grade students in EVD schools were given all of the subtests of the MAT Primary I battery, intended for the grade range 1.5 - 2.4, as well as all of the corresponding subtests for MAT Primary II, the "proper" level for third grade (3.0) students at this point. In this analysis, the Mathematics subtest is excluded, since, as a result of misunderstanding

on the part of the teachers, this subtest was omitted for a large number of students in the sample. These subtests were administered sequentially within subject area, with corresponding subtests from the two levels given 4-6 days apart, with the order of administration of levels within subtests randomly counterbalanced over students in the sample.

It is clearly important that the sample size be maximized; hence, we included the maximum number of students available within the restrictions imposed by the District. Of the total of 801 students eligible for testing, a proportion of valid scores from both levels of corresponding subtests was obtained for 93% or more of the sample. The details of coverage appear in Table (1).

COMPARISON OF CHARACTERISTICS OF SAMPLE AND STANDARDIZATION DATA

Assessments of the "reliability" or, perhaps more properly, generalizability of scores obtained on MAT subtests are provided by the MAT publishers in the form of estimates of coefficients of internal consistency (usually called coefficient alpha or generalized KR20) based upon the data gathered from the 1970 edition's standardization sample; and estimates of the standard error of measurement (SEM) reported are based upon these reliability estimates.[5] In order to assess the degree of comparability of our data with the publisher's data, we estimated these coefficients and SEM's from our own data. The results appear in Tables (2-5).

To summarize these results, we simply observe that in all cases, local consistency coefficient estimates were greater than 0.90; and that, while local estimates were in all cases lower than those reported by the publisher, the modal difference is 0.01, and the greatest difference is 0.02. Estimates of the SEM are equally comparable. Given the comparative homogeneity of the sample, differences of this order of magnitude are surely negligible. On these bases, then, the data from our sample seem wholly comparable to those from the publisher's standardization sample.

In addition, while published reports do not show the first two moments of the distributions of standardization data, the comparability of SEM estimates from our data to those provided by the MAT publisher strongly suggests that the variances from the two samples are reasonably comparable.

Table 1

NUMBER OF STUDENTS WITH SCORES
ON PRIMARY I AND PRIMARY II, BY SUBTEST

Scale	N	% Omitted[a]
Word Knowledge	780	2.62
Word Analysis	776	3.12
Reading	744	7.11
Total	801	

a) For each subtest, the percentage of the total number of students tested for whom at least one subtest score was missing.

Table 2

RELIABILITY COEFFICIENTS[b]

SUBTEST	PRIMARY I		PRIMARY II	
	National[a]	Local	National[a]	Local
Word Knowledge	.94	.932	.95	.940
Word Analysis	.94	.916	.93	.907
Reading	.96	.948	.95	.935

a) Source: MAT Teacher's Handbook.

b) Computed as KR20's: See footnote (5).

Table 3

STANDARD ERRORS OF MEASUREMENT
FOR THREE SUBTESTS: GRADE EQUIVALENTS

SUBTEST	PRIMARY I		PRIMARY II	
	National[1]	Local	National[2]	Local
Word Knowledge	.2	.24	.3	.20
Word Analysis	.2	.26	.3	.30
Reading	.2	.22	.3	.23

1) Fall standardization: grade = 2.1.

2) Spring standardization: grade = 2.7.

Table 4

STANDARD ERRORS OF MEASUREMENT
FOR THREE SUBTESTS: STANDARD SCORES

SUBTEST	PRIMARY I		PRIMARY II	
	National[1]	Local	National[2]	Local
Word Knowledge	2.7	3.2	2.5	2.5
Word Analysis	2.3	2.8	2.8	3.2
Reading	2.5	2.8	2.7	3.2

1) Fall standardization: grade = 2.1.

2) Spring standardization: grade = 2.7.

Table 5

STANDARD ERRORS OF MEASUREMENT
FOR THREE SUBTESTS: RAW SCORES

SUBTEST	PRIMARY I		PRIMARY II	
	National[1]	Local	National[2]	Local
Word Knowledge	1.7	1.8	2.0	2.5
Word Analysis	2.0	2.1	2.0	2.4
Reading	2.2	2.4	2.3	2.8

1) Fall standardization: grade = 2.1.

2) Spring standardization: grade = 2.7.

DESCRIPTION OF ANALYTIC FRAMEWORK

At various points heretofore, we have indicated a concern with the validity or usefulness of the SS scale on both the individual and aggregate levels. Our analytic framework, then, is designed to take account of both of these areas; for convenience, we discuss them seriatim. At this point, we should mention that, while much of the original analysis was reported in a context of comparisons of various linear models; our present discussion, because of time and space limitations, will be restricted, on the individual level, to an assessment of the goodness of fit of correlations obtained from the data, both observed and corrected for putative error of measurement, to the publisher's implicit hypothesis that, in terms of transformed scores (SS), it doesn't matter which test level we use. Note that this paraphrase of Gulliksen's (1950) vernacular definition of parallelism is intentional: if this implicit hypothesis is to receive support, it should be the case that correlations between subtests within domains, when based upon transformed scale scores (SS), should approximate the reported reliabilities of either level; and this, of course, implies that disattenuated correlations should approach 1.00.[6]

Of course, one does not expect complete substitutability across all test levels. For example, a sample of eighth grade students would all be expected to get an essentially perfect score on a test intended for second graders, almost without regard to their scores on a test for eighth graders. For adjacent tests, however, or for tests two steps apart, the tests are designed to permit substitutability, as the publishers themselves have said.

INDIVIDUAL LEVEL ANALYSIS

Although our primary interest is in the performance of the SS, we also included in our analyses parallel assessments of both the Grade Equivalent (GE) scale, since it is, whether merited or not, widely used; and the raw scores, primarily as a baseline. (Inasmuch as the transformation procedure used to relate raw scores to SS assumes that the various raw scores are related by a linear transformation, it may appear that the use of raw scores does not provide a true data baseline, as

against the theoretical baseline discussed above. In the event, however, nothing more suitable is available, we include it here as a matter of interest.)

For each of the subtests included in our study, we report in the following tables, for GE, SS and raw scores, estimates of the observed correlations and correlations corrected for hypothesized error of measurement.

Now, since we discuss in considerable detail elsewhere (Barker and Pelavin, 1975) the propriety of using consistency coefficients as estimates of reliability, we simply briefly outline the argument here. The simplest and perhaps most defensible interpretation of coefficient alpha in the context of test theory is as an index of "behavior domain validity" (Tryon, 1957), i.e., the correlation between scores on a sample from a domain and scores on the total domain. On this interpretation (and derivation: cf, e.g., Kaiser and Michael, 1975), scores in the domain are taken to be "true scores," and, perhaps more to the immediate point, the only source of error of measurement theoretically allowed is error arising from the fact that a domain is sampled instead of exhaustively surveyed in any particular test.

Hence, if estimates of the SEM of a test are to be based (as those of MAT 70 are) on estimates of alpha, the implicit claim is that sampling error alone contributes to error of measurement; and it follows that, if the validity of this claim is to be tested, as here, then estimates of alpha are the correct disattenuation coefficients. (Barker and Pelavin, 1975; Barker, 1975(b)).

Moreover, as should be clear from the description of the design of this study, the interval of time between administration of alternate levels of any subtest is clearly too small for any measurable expected change in actual knowledge to occur. (That is, changes in the relative rank order of the students between test points can hardly be attributed to differential rates of change of learning, since, on the whole, only negligible amounts of learning can be expected to occur.)

Finally, a check of the data for order effects disclosed that no significant effects were present; this finding was confirmed in Pelavin and Barker, (forthcoming; 1976).

A cursory examination of the results presented in Tables (6-8) is sufficient to show that between-level correlations within subtest areas (domains) do not approach at all closely the putative reliability of either level, nor do the disattenuated correlations approach 1.00. Neither do we observe any significant differences in the values of these estimates dependent upon the scale used; in fact, the estimates based upon raw scores and SS are literally indistinguishable. We may summarize these findings by quoting the means over all scales and subtests: RBAR for observed scores = .739; for disattenuated estimates, RBAR = .798.

It is also clear, then that if these estimates from observed data are to be taken as parallel form estimates of reliabilities, the resultant estimates of the SEM must be far higher than those reported by the publisher, than those computed on the basis of estimates of alpha from our own data. Furthermore, given that within-class variances are homogeneous, a hypothesis that cannot be rejected on the basis of our data, the errors of estimate based upon these data (i.e., the pooled standard deviation of residuals about the regression line predicting subtest scores on one level from those on another) must be quite large; in fact, on the whole, they will approximate 2/3 of the standard deviation of obtained scores.[7]

On the whole, then, we must conclude from these data that the validity of the SS scale on the individual level has not been demonstrated. We will return to a discussion of this finding in the penultimate section of this paper, following the presentation of results of the aggregate level analysis.

AGGREGATE LEVEL ANALYSIS

In educational evaluation, the following situation is not at all uncommon: one wishes to assess the amount of cognitive gain of some group of students over, say, a period of one year. Now, as we have pointed out above, it often happens that testing students at both points in time with the same test is inappropriate; this is just the situation of longitudinal comparisons for which, as we have said, score transformations (in this case, the SS transformation) are in part designed. If we assume that the SS on two adjacent levels of some domain (subtest)

Table 6

CORRELATIONS BETWEEN CORRESPONDING SUBTESTS,
GRADE EQUIVALENT UNITS, CORRECTED FOR ATTENUATION

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SUBTEST	N	$R_{I,II}$	$R^2_{I,II}$	$R_{kk,I}$	$R_{kk,II}$	R_{tt}	R^2_{tt}	R^2_{Diff}
Word Knowledge	780	.660	.436	.932	.940	.705	.497	.061
Word Analysis	776	.795	.632	.916	.907	.872	.761	.129
Reading	744	.694	.482	.948	.935	.757	.543	.016
Mean		.716				.778		

(3) = raw score correlation

(4) = raw score common variance

(5), (6) = K-R-20 estimated from present data

(7) = raw score correlation with attenuation correction; estimated "true score" correlation

(8) = estimated "true score" common variance

(9) = estimated common variance increase [(8) - (4)]

Table 7

CORRELATIONS BETWEEN CORRESPONDING SUBTESTS, STANDARD SCORES, CORRECTED FOR ATTENUATION

(1) SUBTEST	(2) N	(3) $R_{I,II}$	(4) $R^2_{I,II}$	(5) $R_{kk,I}$	(6) $R_{kk,II}$	(7) R_{tt}	(8) R^2_{tt}	(9) R^2_{Diff}
Word Knowledge	780	.718	.516	.932	.940	.767	.588	.072
Word Analysis	776	.825	.681	.916	.907	.905	.819	.138
Reading	744	.708	.501	.948	.935	.752	.566	.065
Mean		.750				.808		

(3) = raw score correlation

(4) = raw score common variance

(5), (6) = K-R-20 estimated from present data

(7) = raw score correlation with attenuation correction; estimated "true score" correlation

(8) = estimated "true score" common variance

(9) = estimated common variance increase [(8) - (4)]

Table 8

CORRELATIONS BETWEEN CORRESPONDING SUBTESTS, RAW SCORES, CORRECTED FOR ATTENUATION

SUBTEST	N	$R_{I,II}$	$R^2_{I,II}$	$R_{kk,I}$	$R_{kk,II}$	R_{tt}	R^2_{tt}	R^2_{Diff}
Word Knowledge	780	.687	.472	.932	.940	.734	.539	.067
Word Analysis	776	.786	.618	.916	.907	.862	.744	.126
Reading	744	.778	.605	.948	.935	.826	.683	.078
Mean		.750				.807		

(3) = raw score correlation

(4) = raw score common variance

(5), (6) = K-R-20 estimated from present data

(7) = raw score correlation with attenuation correction; estimated "true score" correlation

(8) = estimated "true score" common variance

(9) = estimated common variance increase [(8) - (4)]

are not *directionally biased*, then it would seem reasonable to compare, say, Spring with Fall scores in transformed score units, and take, say, the mean difference over the group of interest as an unbiased estimate of cognitive gain in the domain.

Now note that, in substituting, as it were, one test level for another, we subject our scores to what Gulliksen (1950) calls error of substitution; however, this simply means that, in addition to the usual error of measurement, we incur some error increment by comparing two presumably independent samples of a domain with each other. It does not, however, entail what we are here calling *directional bias*. The nature of this bias is explicated in detail in Barker and Pelavin (1975); here, we simply provide a brief illustration.

Letting I, II index different test levels (where, in the present case, II = I + 1), and 1, 2 index different times of administration (Fall and Spring, say), we wish to assume that

$$E(II2 - I1) = g, \text{ say,} \tag{1}$$

where E is the expected value operator,

and g denotes "true growth."

Now, we would assume that, in any units,

$$E(II2 - III) = g; \tag{2}$$

i.e., (2) differs from (1) only in that we have substituted I for II at time (1), a procedure which the SS (or equivalent transformation) is designed to permit.

Therefore, writing the identity

we have
$$(II2 - I1) = (II2 - III) + (III - I1),$$
$$E(II2 - I1) = E(II2 - III) + E(III - I1) = g \tag{3}$$

if
$$E(III - I1) = 0. \tag{4}$$

That is, if (4) does not hold, then the comparison (1) is *directionally biased;* in this case, the validity of (4) is testable from the data at hand; and a finding of statistically significant departures of (4) from zero would indicate that any differences found were likely to be reliable. It would then remain to consider how practically significant such differences might be.

Said another way, a finding that (4) does not hold would indicate that, ceteris paribus, it *does* matter which level of a domain test is used, since a comparison of an aggregation of students on *both* levels administered at virtually the same time (1) would show some reliable difference in, say, the mean of their scores, attributable almost solely to the level of the test of the domain which was administered.

The results of aggregate comparisons for all subtests (domains) for both GE and SS units appear in Tables (9-10).

An examination of these tables shows that, for all three GE comparisons and two of three SS comparisons, reliable directional bias appears (for these five, maximum $p = .002$). The fact that, for the vocabulary subtest, $p = .236$ for SS, may be considered an argument in favor of the SS scale as against the GE; however, the other, evidently reliable differences, are quite large, when we consider that they constitute means over a minimum of 744 students.

Now, the magnitudes of these differences in SS have no prima facie interpretation, although they may be given a rough interpretation in terms of percentiles (see Barker and Pelavin, 1975). However, the closely related GE differences may be given a rough interpretation in the following way. If we assume that .1 GE is the equivalent of about one month's gain for the average student, the differences we find are the equivalent of 10% - 20% of an expected year's growth, although these differences appear from tests administered over a period of less than one week.

On the other hand, these are not, in the usual parlance, average students; for students of the sort at hand, the usual estimate of expected growth over a school year is about 0.6 GE (cf, e.g., Fennessey, 1973). If we adopt this rough expectation, the size of the directional

Table 9

TEST OF MEAN DIFFERENCES PRIMARY I AND PRIMARY II GRADE EQUIVALENTS, WITHIN CORRESPONDING SUBTESTS

SUBTEST	MEAN DIFFERENCE (I-II)	SD	SE	DF	t	P	95% CONFIDENCE LIMITS UPPER	LOWER
Word Knowledge	.082	.721	.026	779	3.155	.002	.133	.031
Word Analysis	-.183	.602	.022	775	8.442	<.001	-.140	-.226
Reading	.084	.721	.026	743	3.188	.002	.135	.033

Table 10

TEST OF MEAN DIFFERENCES PRIMARY I AND PRIMARY II
STANDARD SCORES, WITHIN CORRESPONDING SUBTESTS

SUBTEST	MEAN DIFFERENCE (I-II)	SD	SE	DF	t	P	95% CONFIDENCE LIMITS	
							UPPER	LOWER
ord Knowledge	0.364	8.564	0.307	779	1.187	.236	0.966	-0.238
ord Analysis	-2.115	6.106	0.219	775	-9.642	.001	-1.686	-2.554
eading	1.497	9.307	0.341	743	4.385	.001	2.165	0.829

bias here is in the range 16% - 33% of a year's growth: very large proportions indeed, if we consider that they appear to be functions of the system of measurement itself, not of real gains; and that effects of educational "treatments" are not usually so large (Averch, et al., 1970; Levin, 1970; Crain, 1973; Acland, et al., 1975; e.g.).

That is, if proportional gains of this order of magnitude relative to expectation were found over the course of a year for educational interventions, and they were assumed to be unbiased (an untestable assumption, under most circumstances--which is, of course, why we are testing it here); depending upon the direction of the bias, the interventions in question might be judged either rather sensational successes or fairly disappointing.

And this, of course, is precisely how cognitive scores are likely to be used by evaluators and/or policymakers; hence, the finding of reliability biases of this magnitude is fairly disturbing.

Now, of course, in a true randomized experimental design, systematic bias of this sort would not bias estimates of treatment effects, ceteris paribus; unfortunately, most educational evaluation cannot claim even incomparable comparison groups (for which bias of this sort could make a difference); the paradigm of comparing observed with expected differences is far more common; and it is for just such comparisons that bias of the kind here discussed and illustrated is confounded with "true" growth.

As discussed more fully in Barker and Pelavin (1975), we find no reason to attribute these biases to floor and/or ceiling effects, and, in any case, if we were to make such an attribution, the finding that the differences are *not* always in the same direction would seem to invalidate the attribution.

In short, of the two classes of findings presented here, those just discussed seem to us to be potentially the more serious. If we consider the task of the evaluator and/or policymaker to be one of binary classification of educational treatments into what we might call go/no go categories, then we can see that it is as true of treatments as it is of persons that, if scores are to be used for classification, they must

be quite highly reliable. But this is just what we have found the scores under investigation *not* to be.

DISCUSSION

Alternative explanations for the findings presented here are exhaustively discussed in Barker and Pelavin (1975); but the most likely conclusion may be more briefly expressed here. It is simply that, when we realize that the measurement system which we are discussing here was not subjected by the publisher to any known validation of the sort here reported--granting at once that this is, as we all know, no simple thing to do; the likeliest explanation for the results found (and replicated: Pelavin and Barker, forthcoming, 1976) is simply the invalidity of the basic metric of the system.

For example, it can be shown (e.g., Barker, 1975(b)) that, given the assumptions underlying the estimate of the SEM by the publisher, disattenuated between-level correlations are a function of the mean item covariances within each level and between levels; that, in fact,

$$R(T_x,T_y)^2 = \bar{C}_{ip}^{\,2}/\bar{C}_{ij}\,\bar{C}_{pq}, \qquad (5)$$

where T_w denotes disattenuated scores on test (w),

$\bar{C}_{ip}^2$ = mean squared item covariances between levels,

and $\bar{C}_{ij}$, $\bar{C}_{pq}$ = mean item covariances within each level.

If all of these mean covariances are not roughly equal to the extent that the mean between covariance is less than the geometric mean of the mean within covariances, the between-level disattenuated correlation will be less than 1.00. However, this would indicate, on the assumptions underlying reported estimates of SEM, that rather different domains were sampled: for tests of this sort, this seems generally unlikely.

We should add that, while, to our knowledge, the studies reported here and in Pelavin and Barker, (forthcoming; 1976) are the only ones extant that set out systematically and specifically to test the validity

of system metrics, it is not the only one which suggests directional bias (see, e.g., Ayrer and McNamara, 1973).

In short, we cannot conclude with any confidence that the instruments herein assessed provide dependable bases either for individual student assessment, or program evaluation as usually performed.

CONCLUSION

At this point, given current practice in evaluation and assessment, it seems natural to ask, What is the practical import of these findings? In short, what are we to do?

Unhappily, we cannot, on the basis of these analyses, present any clear answer to this question; we may, however, present some suggestions.

One of these is thought to be rather difficult (by some, impossible) to implement. We feel that the difficulty is exaggerated, but that does not alter the feelings of those who are responsible for evaluation. It is simply that many more evaluations than at present be designed as randomized true experiments, rather than the quasi- or non-experiments that are the rule today. This would at least enable us to have a bit more confidence that estimates of treatment effects were, in truth, the unbiased estimates which we must usually, when assessing the treatments, at least implicitly assume that they are.

Secondly, we would suggest that the sole dependence which we so often find upon scores from SATs be rather radically changed. At the very least, it would seem the better part of wisdom to administer more than one battery of such tests, time consuming though that may be; if the results from multiple administrations are *not* convergent, caution in interpretation is of course indicated. (Note, however, that convergence is *not* proof of validity of either or both sets of scores.) Even better, it seems to us, would be the additional administration of tests specifically designed to measure learning of just what is taught. Not that this is easy to do, either: quite the contrary. But one must recall that SATs are not validated in the sense in which we usually think of validation, i.e., against an explicit criterion; rather, as Goslin (1967), (among others), points out, SATs are themselves in a very real sense taken as *criterial*. But, for that very reason, given that they are designed

for extremely wide usage, they are truly criterial for few if any real existent programs or curricula.

Despite the argument made above that floor/ceiling effects do not account well for the present findings, it is expectable in general, and true in this case, that the distribution of raw scores on the lower level test is somewhat skewed relative to the upper. Since the scaling method used to map raw scores onto a common (Standard Score) scale requires, for validity, only that the two sets of raw scores be related by a linear transformation, clearly the method takes no account of the third moments of the distributions. (See also Gulliksen, 1950.) While the relative skew in these data is not large (Barker and Pelavin, 1975), it is arguable that failure to correct for even a small relative skew could invalidate the scale.

Two remedies suggest themselves for this situation, apart from developing a method which does take account of third moments. One of these amounts to decreasing the number of levels of the test, while including a certain amount of overlap; indeed, there is informal evidence that this ameliorates the problem. The other would involve administration of a careful pretest, so that individual students would be administered the level on which they would be most likely to achieve a score in the middle of the range. Research is underway to assess the usefulness of this strategem; however, data presented in detail in Barker and Pelavin (1975), comparing between-level score differences for students grouped into quartiles on one level, suggests that rather large differences remain even for students relatively near the center of the distributions on both levels.

Again, if dependence is to be placed upon test scores, for a program of any scope or importance, it might well be necessary, if SATs are to be used, for the evaluators to undertake extensive and rigorous metric validation and, if required, reconstruction prior to beginning the evaluation. Now, this is no doubt a difficult and expensive undertaking, but neither so difficult nor so expensive as developing and fielding the programs which are to be evaluated. If this greater sum is not to be placed at hazard by relatively unreliable and invalid assessment criteria--an unthinkable, but nonetheless widespread phenomenon--one can only, we believe, conclude that these difficulties and expenses must be conquered and paid. We have tried to set out some of the ways in which this might be done.

NOTES

1. This research was undertaken while the first author was associated with The Rand Corporation, and was supported by the National Institute of Education. The authors would like to express their deep gratitude for the invaluable advice and assistance of Dr. T. S. Donaldson and Carol N. Frost, both then at The Rand Corporation; to Professor Ward Keesling, UCLA; and Mr. David P. Mandel, NIE.
2. We should make it very clear that this research is not intended specifically to criticize the MAT; on the contrary, the MAT was chosen precisely because it has been found (Hoepfner, et al., 1970) to be exemplary of the genre. It is our belief that the findings reported here are probably applicable to most, if not all, of the SATs in wide use.
3. It seems likely that this state of affairs is in no small part a result of the seemingly irresistible pressure upon both sponsors and evaluators to attempt to measure cognitive growth or status even when that is not the sole or even primary aim of the program. This, in turn, is probably because, in fact, it is widely felt that cognitive outcomes *are* important (which is probably true) *and* that they, among the range of possible outcomes, are unusually easy to measure (which is extremely doubtful).
4. However, following the outcome of the study herein reported, the issue was deemed of sufficient importance to mount a much wider study, which was in part a replication of this one. These results are reported in Pelavin and Barker (1975; 1976); and they do, in fact, support the results and conclusions reported here.
5. In fact, the publisher's estimates are based upon one of the Saúpe (1961) estimates of KR20; however, the differences observed in practice between the estimates based upon the two procedures are entirely negligible.
6. For a detailed discussion of this matter, see Barker and Pelavin (1975); and Barker (1975(b)).

7. That is, if we let

V = variance of obtained scores,

V_r = variance of residuals,

then $V_r = V(1 - r_{12}^2)$,

or $SD_r = \sqrt{V}\sqrt{(1 - .546)} = .674\sqrt{V}$.

REFERENCES

Acland, H., Barker, P., Crain, R. L., Pelavin, S. H., and Sitgreaves, R., *Investigation of the Impact of the Emergency School Assistance Program on Black, Male, 10th Grade Achievement*, Santa Monica: The Rand Corporation, 1975, unpublished.

Angoff, W. H., Scales norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*, Second Edition. Washington: American Council on Education, 1971.

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., and Pincus, J., *How Effective is Schooling?* Englewood Cliffs, New Jersey: Educational Technology Publications, 1974.

Ayrer, J. E. and McNamara, T. C., Survey testing on an out of level basis, *Journal of Educational Measurement*, 1973, *10*, 79-83.

Barker, P., *Analytic Approaches to Longitudinal Race and Class Indicators*, Santa Monica: The Rand Corporation, 1974, unpublished.

Barker, P., *Issues in Measuring Student Cognitive Outcomes in Education Interventions: The Case of Alum Rock*, Santa Monica: The Rand Corporation, 1975(a), unpublished.

Barker, P., Test reliability and the correction for attenuation, 1975(b), (in review).

Barker, P. and Pelavin, S. H., *Concerning Scores and Scale Transformations in Standardized Achievement Tests, Their Accuracy and Dependability for Individual and Aggregation: The Case of MAT 70*, Santa Monica: The Rand Corporation, 1975, unpublished.

Crain, R. L., *Southern Schools*, Volumes I and II, Chicago: NORC, 1973.

Fennessey, J., *Using Achievement Growth to Analyze Educational Programs*, Baltimore: Johns Hopkins University, Center for Social Organization of Schools, 1973.

Goslin, D. A., *Criticism of Standardized Tests and Testing*, Princeton: CEEB/ETS, 1967.

Guilford, J. P., *Psychometric Methods*, Second Edition, New York: McGraw-Hill, 1954.

Gulliksen, H., *Theory of Mental Tests*, New York: Wileym, 1950.

Harcourt Brace Jovanovich, Inc., *Development of the Standard Score System for the 1970 Edition of MAT*, New York: Author, 1972.

Harcourt Brace Jovanovich, Inc., *Development and Use of the Grade Equivalent Scale*, New York: Author, 1973.

Hoepfner, R., et al., *CSE Elementary School Test Evaluations*, Los Angeles: UCLA, Center for the Study of Evaluation, 1970.

Jencks, C. and staff, *Education Vouchers: A Report on Financing Elementary Education by Grants to Parents*, Cambridge: Center for the Study of Public Policy, 1970.

Kaiser, H. F. and Michael, W. B., Domain validity and generalizability, *Educational and Psychological Measurement*, 1975, *35*, pp. 31-36.

Levin, H., A new model of school effectiveness, in U.S. Department of HEW, *Do Teachers Make a Difference?* Washington: USGPO, 1970.

Pelavin, S. and Barker, P., *An Investigation of the Generalizability of Scaled Scores in MAT 70*, Santa Monica: The Rand Corporation, forthcoming.

Pelavin, S. and Barker, P., *A study of the generalizability of standardized achievement tests*, presented at the Annual Meeting of the American Educational Research Association, 1976. Santa Monica: The Rand Corporation, P-5678.

Porter, A. C. and Chibucos, R. R., Common problems of design and analysis in evaluative research, *Sociological Methods and Research*, 1975, *3*, pp. 235-257.

Saupe, J. L., Some useful estimates of the Kuder-Richardson formula number 20 reliability coefficient, *Educational and Psychological Measurement*, 1961, *21*, pp. 63-71.

Thurstone, L. L., A method of scaling psychological and educational tests, *Journal of Educational Psychology*, 1925, *16*, pp. 433-451.

Tryon, R. C., Reliability and behavior domain validity: Reformulation and historical critique, *Psychological Bulletin*, 1957, *54*, pp. 229-249.

Weiler, D., et al., *A Public School Voucher Demonstration: The First Year at Alum Rock*, Santa Monica: The Rand Corporation, 1974, R-1495-NIE.

Weiner, S. S. and Kellen, K., *The Politics and Administration of the Voucher Demonstration in Alum Rock: The First Year*, Santa Monica: The Rand Corporation, 1974, unpublished.